Method And Apparatus For Structuring Texts

[0001] The present application hereby claims priority under 35 U.S.C. §119 on German patent application number DE 102 45 876.6 filed September 30, 2002, the entire contents of which are hereby incorporated herein by reference.

Field of the Invention

[0002] The invention generally relates to a method and apparatus for converting unstructured text information into a structured format.

Background of the Invention

[0003] Particularly in medical engineering, many free text reports are produced today which are recorded in the computer using dictaphones and/or voice recognition technologies, for example. The problem when handling these reports is that automatic access to small information parts, "atomic information", is almost impossible because the content contains no or just a very coarse structure. Free text reports are therefore very unsuitable for structured presentation and evaluation of the information.

[0004] In such free text reports, only integrated information is processed. This information cannot be used for automatic evaluations. Thus, the information it contains is thus lost for this purpose. This problem is growing as the need for access to the atomic information, for example for the purpose of coding, increases.

[0005] Aho, Alfred V. et al, "Compilers - Principles, Techniques and Tools", Addison Wesley, Reading, Massachusetts, 1986, pages 4 to 11, the entire contents

of which are incorporated herein by reference, describes the principle of parsing.

[0006] Wormek A.K. et al., "SAM: Speech-Aware Applications in Medicine to Support Structured Data Entry" ,the entire contents of which are incorporated herein by reference, discloses a method for the structured input of data by voice.

[0007] In these documents, unstructured text information is converted into a structure on the basis of the derivation of one structure from another. These resultant structures also cannot be used for automatic evaluations.

SUMMARY OF THE INVENTION

[0008] An embodiment of the invention is based on an object of providing a method and an apparatus which allow simple, automated conversion of unstructured text information from free text reports into a structured, evaluatable format.

[0009] An embodiment of the invention achieves an object via a method having the following steps:

a) structuring rules for structuring the unstructured text information are input,
b) unstructured text information is recorded,
c) the unstructured text information is parsed in order to produce small text fragments,
d) text units of the unstructured text information are searched for text fragments defined in the structuring rules,
e) the text fragments of the unstructured text information are structured on the basis of conditions stipulated in the structuring rules.

[0010] The structuring rules to be defined parse the free text report, i.e. break it down into smaller units, and convert it into a structure which allows a program to evaluate this information. Such a rule contains information relating to the text fragments for which the free text report needs to be searched, which structure element is represented thereby, and additional information about how the structure needs to be set up.

[0011] In line with the invention, unstructured text information can be recorded in step b) by a microphone, with a voice recognition program being used for conversion into unstructured text information.

[0012] Advantageously, the structuring rules can contain information relating to the text fragments for which the free text report needs to be searched, about which structure element is represented thereby and about how the structure needs to be set up.

[0013] An embodiment of the invention achieves an object for the apparatus by way of an input apparatus for unstructured text information, an input apparatus and a memory apparatus for structuring rules, an extraction apparatus for small text units from the unstructured text information, a structuring apparatus for producing structured text information on the basic of the structuring rules, and an evaluation apparatus for the text units in the structured text information.

[0014] Evaluatable unstructured text information can be input directly if the input apparatus for unstructured text information has an associated apparatus for voice recognition.

[0015] It has been found to be advantageous if DICOM-SR or XML is used as structured format for the structured text information.

BRIEF DESCRIPTION OF THE DRAWINGS

[0016] The present invention will become more fully understood from the detailed description of preferred embodiments given hereinbelow and the accompanying drawings, which are given by way of illustration only and thus are not limitative of the present invention, and wherein:

Figure 1 shows an apparatus in accordance with an embodiment of the invention for structuring texts, and

Figure 2 shows a method in accordance with an embodiment of the invention for structuring texts.

DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENTS

[0017] Figure 1 shows an apparatus in accordance with an embodiment of the invention for structuring texts. The apparatus can be implemented in a personal computer (PC), for example. A keyboard 1, for example, may be used for inputting structuring rules and possibly free text reports. In addition, the apparatus can have a voice input apparatus 2, for example a microphone or a cassette player, which can be used to input the free text reports into the PC. The voice input apparatus 2 has an apparatus 3 for voice recognition, for example with a voice recognition program, connected to it which can be used to convert the spoken free text reports into text information.

[0018] The keyboard 1 is connected to a memory apparatus 4 for structuring rules and to a memory apparatus 5 for text information, to which the apparatus 3 for voice recognition is also connected. The memory apparatus 5 for text information has an extraction apparatus 6 connected to it which recognizes and identifies small text units from the unstructured text information. The extraction apparatus 6 and the memory apparatus 4 for the structuring rules have a structuring apparatus 7 for producing structured text information connected to them which converts the extracted text units into a structured format on the basis of the stipulated and stored structuring rules. The structuring apparatus 7 has an evaluation apparatus 8 connected to it which allows a check for small, structured text units for further evaluation.

[0019] In a medical facility, free text reports are recorded, for example using a dictaphone, and are later transferred to the computer by a secretary using a writing program via the keyboard 1. A free text report can also be converted into a written text by the apparatus 3 for voice recognition, using an appropriate voice recognition program, the free text report being able to be input directly into a personal computer by means of dictation or subsequently using a player for dictation cassettes.

[0020] To allow later evaluations of the stocks of data produced in this manner, the free text reports are converted into a structured format, for example DICOM-SR or XML, in addition to their original format. For this purpose, rules are defined which stipulate the systematics of conversion.

[0021] The starting point is unstructured text information 9, shown in Figure 2, which has been produced by way of dictation or free text input. This text information 9 is used as input for an apparatus which is intended to convert this unstructured text information 9 into a structured form.

[0022] Figure 2 gives the following as an example of unstructured text information 9:

> Indication: Diaphoresis. Rule out abnormalities of regional wall movements. Check hypertonic cardiomyopathy. Rule out myocardial infarction. Assess the left of the sputum component from the left ventricle. Rule out an aneurysm of the left ventricle. History: other relevant histories include: further cocaine abuse. Previous CV procedures:
> Studyinfo. The study was carried out under general anesthesia.

[0023] To convert this unstructured text information 9 into a structured form, structuring rules 10 are input into this apparatus using the keyboard 1 and are stored in the memory apparatus 4, these structuring rules forming the basis of the conversion.

[0024] These structuring rules 10 define those text fragments for which the text needs to be searched and what result the finding of such a text fragment has in the conversion. In the example described below, finding the text fragment "Indication", for example, signifies that a new element which describes an indication is inserted into the structure.

[0025] The text below gives examples of such structuring rules 10, which are shown in Figure 2. The general basis is that structuring rules 10 are defined which

stipulate, on the basis of the finding of text fragments, how unstructured text information 9 is transferred to a structured form.

[0026] If the text contains the word "Indication", then the word needs to be handled with open actions under element "Indication". The same applies for the word "History" as "History" element and for "Studyinfo" as "Studyinfo" element.

[0027] If the text contains the word "Diaphoresis", then it needs to be inserted as an action under element "Indication". The word "Cocaine abuse" in the text needs to be inserted under element "History entry". The term "General anesthesia" needs to be inserted under element "Studyinfo".

[0028] These and other structuring rules 10 which have been input once, but can be changed at any time, are used to put unstructured text information 9 from the free text report into a structured form, so that the structured text information 11 which has now been obtained and which is described below can be searched for particular terms.

<Report>
<Indications>
<Indication> Diaphoresis</ Indication >. Rule out abnormalities of regional wall movements. Check hypertonic cardiomyopathy. Rule out myocardial infarction. Assess the left of the sputum component from the left ventricle. Rule out an aneurysm of the left ventricle.
</Indications>
<History>
Other relevant histories include: further <History entry> Cocaine abuse <History entry>. Previous CV procedure(s):

</History>
< Studyinfo >
The study was carried out under <Studyinfo> general anesthesia <Studyinfo>.
</Studyinfo>
</Report>

[0029] In this case, the invention involves unstructured text information being converted into a structure on the basis of the rule-based interpretation of contents.

[0030] Thus, by way of example, two documents can contain the following text passages:

a) "The patient was subjected to an extensive examination. An intestinal tumor was diagnosed."

b) "Following a CT-based examination, a tumor in the intestinal tract was diagnosed".

[0031] To structure the diagnosis, the following rules can be applied:

1. If a sentence contains the words "diagnosed", "diagnostic result" or "diagnosis", then it contains information relating to diagnosis.

1.1. If the same sentence contains the word "tumor" or "malignant tumor", a tumor has been discovered.

1.1.1. If the same sentence contains the word "intestine" or "intestinal tract", then intestinal cancer has been diagnosed.

1.2. If the sentence contains the word "intestinal tumor" or "intestinal cancer", then intestinal cancer has been diagnosed.

[0032] The same text fragment is analyzed in this manner from a wide variety of aspects. The knowledge obtained from these analyses is then converted into corresponding structures:

<Diagnosis>
<Code> DF-0044A </CODE>
<Meaning> Intestinal cancer </Meaning>
</Diagnosis>

[0033] It is thus possible to access atomic information automatically, since the content is given a finely structured form by the inventive apparatus. Hence, free text reports can also be used for structured presentation and automatic evaluation of the information.

[0034] Exemplary embodiments being thus described, it will be obvious that the same may be varied in many ways. Such variations are not to be regarded as a departure from the spirit and scope of the present invention, and all such modifications as would be obvious to one skilled in the art are intended to be included within the scope of the following claims.